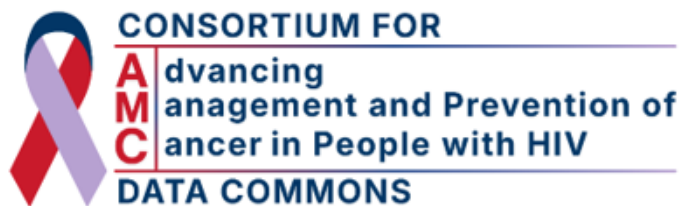

**Consortium for Advancing Management and Prevention of Cancer in
People with HIV Data Commons (AMC-DC)**

Data Standardization Plan for PCR / Sequencing



Template v2.0; 07APR2021

Assay Plan v2.0; 14OCT2026

Table of Contents

1. Glossary	3
2. Introduction.....	3
3. Raw Data Files and Format	4
3.1 Raw Data Files.....	4
3.2 Raw Data Format Description.....	4
4. Laboratory Data Pre-Processing	5
4.1 Data Pre-Processing Steps	5
4.2 Final Laboratory-Processed Data Format	6
5. Experimental Information and Key Meta Data Descriptors	10
6. Quality Metrics.....	14
7. Manifest of File Collection to be Submitted to the ODMC	17
8. ODMC Data Processing	18
9. Approval	19

1. Glossary

Data Formats: A collection of requirements / templates to ensure that the data adheres to a certain standard including minimal required data elements, accepted values, control vocabulary and/or ontologies to be used for certain fields.

Meta data: Collection of data to be collected about the data. This includes information about each field to be collected for a certain assay, experimental information such as instrument, instrument configuration, measurement unit/description, descriptions of the file collection, and other information that is valuable for analysis. Minimal meta data descriptors are a subset of meta data that are considered to be essential.

Operations and Data Management Center (ODMC): provides operational, logistical, communications, contracting, protocol development, site management, data and specimen management, monitoring, regulatory and medical monitoring support to the AMC.

Statistics and Data Analysis Center (SDAC): provides statistical design and analysis of AMC studies

AMC Network Research Laboratories (NRLs): provide expertise in virology, viral oncology (e.g., KSHV, HPV, EBV), pharmacology, pathology, hematopathology, genomics, and advanced HIV diagnostics for correlative studies in support of AMC clinical trials

Data Standardization Teams: Develop Data Standardization Plans for each assay type specifying expected data pre-processing, sample quality metrics, data formats, data processing workflows, and experimental meta data descriptors.

2. Introduction

This document outlines specifications for raw data files/formats, data pre-processing, sample quality metrics, and experimental meta data descriptors for the following assays:

PCR Assays

PCR
qPCR
RT-PCR
qRT-PCR
Atila
HIV Reservoir
Cepheid GeneXpert
Digital Droplet PCR (ddPCR)

Sequencing Assays

RNA-Seq
DNA-Seq
WGS
WES
DNA targeted (panels)
16S metagenomic sequencing
Clonal immunoglobulin sequencing
Methylation sequencing

Data submitted to the ODMC must follow the standards included in this document. This includes laboratory data generated by AMC network research laboratories (NRLs) and affiliate laboratories.

3. Raw Data Files and Format

3.1 Raw Data Files

Table 1: Raw Data File Overview

Assay(s)	File Extension/Type	Description of File Content	Reference
PCR Assays	n/a	PCR data received as processed files, see Table 5 for details.	n/a
RNA-Seq, WGS, WES, Viral Sequencing, Clonal immunoglobulin sequencing, Metagenomics sequencing, DNA-Seq and DNA targeted (panels)	fastq.gz	gzipped compressed sequence data, including quality scores	https://en.wikipedia.org/wiki/FASTQ_format
RNA-Seq, WGS, WES, Viral Sequencing, Clonal immunoglobulin sequencing, DNA-Seq and DNA targeted (panels)	bam and bai	compressed reference alignment format and associated index file	https://samtools.github.io/hts-specs/SAMv1.pdf

3.2 Raw Data Format Description

For details of the FASTQ and BAM file formats see References in Table 1 above.

There are no raw data elements expected to be provided to the AMC DC for PCR assays. See section 4.2 for processed data elements.

4. Laboratory Data Pre-Processing

4.1 Data Pre-Processing Steps

The following data pre-processing steps will be performed for sequencing assays noted below as applicable (data pre-processing not applicable for PCR assays). Specific software and database versions will be reported to the AMC-DC as outlined in Table 3:

Table 2: Data Pre-Processing Overview (RNA-Seq, WGS, WES, DNA-Seq and DNA targeted (panels))

Step ID	Process Description	Software	Command Line Arguments	Input File Type	Description of Input File Content	Output File Type	Description of Output File Content
1	Read quality trimming	Trimmomatic	java -jar trimmomatic-0.38.jar <PE SE> -threads <n_cores> <in_file_fastq> <out_file_trimmed_fastq> SLIDINGWINDOW:4: <quality_cutoff>	FASTQ file	https://en.wikipedia.org/wiki/FASTQ_format	FASTQ file	https://en.wikipedia.org/wiki/FASTQ_format
2	Reference alignment	HISAT2	hisat2 -p <n_cores> -x index/hisat2_genome_index <in_file_fastq> <out_unmapped_fastq> --known-splicesite-infile <in_file_splice_sites> -S <out_file_sam>	FASTQ file, ENSEMBL GTF file, genome/transcriptome reference sequence	https://en.wikipedia.org/wiki/FASTQ_format , https://uswest.ensembl.org/info/website/upload/gff.html	BAM file	https://samtools.github.io/hts-specs/SAMv1.pdf
3	Gene Expression quantification	featureCounts (Subread package)	featureCounts -T <n_cores> -a <in_file_gtf> -o <out_file_feature_counts> <in_file_bam>	BAM file	https://samtools.github.io/hts-specs/SAMv1.pdf	Features counts.txt	

In the above examples, <> should be populated with the appropriate file/option; | = or; PE = paired end; SE=single end.

Data processing specifications will be reported in the following format. Requirements for collection are noted in the last column titled “Collection Requirement”. Data processing specifications will be submitted in a .csv file to the ODMC.

- Filename will be [Protocol ID]_[assay]_[date]_workflow.csv
- The data must have the columns listed exactly as in the “Variable Name” column in the table below. Included values must be included as listed in the “Validation Rules” column in the table below.

Table 3: Variables to be Submitted in [Protocol ID] [assay] [date] workflow.csv

Variable Name	Type	Length	Description	Reference	Validation Rules	Collection Requirement
Protocol	Num	5	Protocol ID	--	Must match ODMC-specified value for the study	Required
ENSEMBL version	Char	25	ENSEMBL database version	https://uswest.ensembl.org/index.html	Must be recorded in the following format: Version XX Month Year E.g. Version 92 April 2018	Required
Trimmomatic version	Num	4,2	Trimmomatic software version	http://www.usadellab.org/cms/?page=trimmomatic	Must be recorded in the following format: X.XX e.g. 0.36 for Version 0.36	Required
UMI-aware trimming	Char	1	Indicate if UMI-aware trimming performed		Must be recorded in the following format: X e.g. 0 for ‘No’, 1 for ‘Yes’	Required
HISAT2 version	Num	5,1,1	HISAT2 software version	https://github.com/DaehwanKimLab/hisat2	Must be recorded in the following format: X.X.X e.g. 2.1.0 for Version 2.1.0	Required
VCF Multiplexer	Char	25	VCF Multiplexer			Required

4.2 Final Laboratory-Processed Data Format

Data will be submitted in the following processed data format. For sequencing assays, processed data will be submitted in the standard formats outlined in Table 4.

Table 4: Processed Data File Overview for sequencing assays

Assay	File Extension/Type	Description of File Content	Reference
RNA-Seq, WGS, WES, Viral Sequencing, Clonal immunoglobulin sequencing	VCF Aligned read file	Stores gene sequence variations	https://en.wikipedia.org/wiki/Variant_Call_Format
RNA-Seq	gene count matrix	Stores the read counts for each sample in a single file	

Laboratory-processed PCR assay data will be submitted in a .csv file to the ODMC as outlined in Table 5. Requirements for collection are noted in the last column titled “Collection Requirement”.

- Filename will be [Protocol ID]_[assay]_[date]_exp_data.csv
- The PCR data must have the columns listed exactly as in the “Variable Name” column in Table 5 below. Included values must be included as listed in the “Controlled Vocabulary” and “Validation Rules” columns in Table 5 below.

Table 5: Variables to be Submitted in [Protocol ID]_[assay]_[date]_exp_data.csv for PCR assays

Variable Name	Type	Length	Description	Controlled Vocabulary	Validation Rules	Collection Requirement
Protocol	Num	5	Protocol ID	--	Must match ODMC-specified value for the study	Required

Variable Name	Type	Length	Description	Controlled Vocabulary	Validation Rules	Collection Requirement
Specimen ID	Num	9	Specimen number, refers to the GlobalTrace barcode number printed on each cryovial	--	Must be 9 characters long. If a Specimen is being entered as a reference (control) sample, the SN will be labeled as [specified value]	Required
Test Date	Date	8	Date of test	--	xx/xx/xxxx (mm/dd/yyyy)	Required
Assay	Char	4	Assay name	PCR, RT-PCR, qPCR, qRT-PCR, ddPCR, Xpert HPV	Must match one of the values listed in the Controlled Vocabulary column	Required
Assay Descriptor	Char	10	Details of conducted assay	For Xpert HPV, expected values include: HPV16, HPV18_45, P3, P4, P5, SAC	Must match one of the values listed in the Controlled Vocabulary column	Required for Xpert HPV
[Result A Numeric]	Num	13,4	Assay result 1. For viral loads, report copy number. For Xpert HPV, report Ct	--	--	Required for all assays
[Result B Numeric]	Num	13,4	Assay result 2, as applicable.	--	--	Permitted
[Result C Numeric]	Num	13,4	Assay result 3, as applicable.	--	--	Permitted
[Result D Numeric]	Num	13,4	Assay result 4, as applicable	--	--	Permitted

Variable Name	Type	Length	Description	Controlled Vocabulary	Validation Rules	Collection Requirement
[Result A Character]	Char	25	Assay result 1. For viral loads and Xpert HPV, report “negative” or “positive”			Permitted
[Result B Character]	Char	25	Assay result 2, as applicable			Permitted
[Result C Character]	Char	25	Assay result 3, as applicable			Permitted
[Result D Character]	Char	25	Assay result 4, as applicable			Permitted
PLATE	Num	2	Plate number / identifier	--	--	Required for PCR
ROW	Num	2	Row number / identifier	--	--	Required for PCR
COLUMN	Num	2	Column number / identifier	--	--	Required for PCR
WELL	Num	3	Well number / identifier	--	--	Required for PCR
PRIMER	Char	25	PCR primer used	--	--	Required for PCR
Comments	Char	200	Result Comment Text or Cancellation Reason, where applicable	--	--	Permitted

Specimen and data file linkages for sequencing assays will be reported in the following format. Requirements for collection are noted in the last column titled “Collection Requirement”. Specimen linkages will be submitted in a .csv file to the ODMC.

- Filename will be [Protocol ID]_[assay]_[date]_specimen_linkage.csv
- The data must have the columns listed exactly as in the “Variable Name” column in the table below. Included values must be included as listed in the “Controlled Vocabulary” and “Validation Rules” columns in the table below.
- Specimen linkages for PCR assays are defined within processed data file, and therefore not described here.

Table 6: Variables to be Submitted in [Protocol ID] [assay] [date] specimen linkage.csv

Variable Name	Type	Length	Description	Controlled Vocabulary	Validation Rules	Collection Requirement
Specimen ID	Num	9	Specimen number, refers to the GlobalTrace barcode number printed on each cryovial	--	Must be 9 characters long. If a Specimen is being entered as a reference (control) sample, the SN will be labeled as [specified value]	Required for sequencing assays
File Name	Char	256	Name of the file	See details in each section	<i>See details in each section</i>	Required for sequencing assays
File Type	Char	256	Type of the file	Raw Data, Interim-Processed Data, Final-Processed Data	Must match one of the values listed in the Controlled Vocabulary column	Required for sequencing assays

5. Experimental Information and Key Meta Data Descriptors

A summary of the experiment will be provided in a text file with comparable details similar to the example below:

Blood will be collected into 8 mL sodium citrate CPT tubes and then centrifuged at 1800g at room temperature (RT) for 30 minutes (with brakes off). The mononuclear cell layer will be transferred to a 15 mL conical tube and the volume brought to a total of 13 mL with PBS. The PBMCs will then be centrifuged at 300g for 10 minutes at RT, and 2.5 x 10⁶ of the cells will then be transferred to a 2 mL tube. The cells will be centrifuged at 2000g for 5 minutes, the supernatant removed, and the cell pellet resuspended in 350 mL of Buffer RLT (Qiagen) containing 1% b-mercaptoethanol. The samples will then be vortexed to lyse the cells, and then stored at -70°C for later extraction. RNA will be isolated from each sample using the RNeasy Mini kit (Qiagen) with on-column DNase digestion [1]. RNA quality will be assessed using an Agilent Bioanalyzer and 500 nanograms of total RNA will be used as input for library preparation using the TruSeq mRNA Stranded Kit (Illumina) according to the manufacturer's instructions. As an internal control, an ERCC ExFold RNA Spike-In Mix (ThermoFisher) will be also added to each sample well according to manufacturer instructions. Libraries will be validated by capillary electrophoresis on an Agilent 4200 TapeStation, pooled at equimolar concentrations, and sequenced on an Illumina HiSeq3000 [5] at 100SR, yielding 25-30 million reads per sample.

Experimental meta data descriptors will be collected and reported in the following format. Requirements for collection are noted in the last column titled "Collection Requirement". Experimental meta data descriptors will be submitted in a .csv file to the ODMC.

- Filename will be [Protocol ID]_[assay]_[date]_exp_metadata.csv

- The data must have the columns listed exactly as in the “Variable Name” column. Included values must be included as listed in the “Controlled Vocabulary” and “Validation Rules” columns in the table below.

Table 7: Variables to be Submitted in [Protocol ID] [assay] [date] exp metadata.csv

Variable Name	Type	Length	Description	Controlled Vocabulary	Existing Standard/Link	Validation Rules	Collection Requirement
Protocol	Num	5	Protocol	--	--	Must match ODMC-specified value for the study	Required
Assay Type	Char	50	Assay type Name	PCR/Sequencing	--	Must match the value listed in the Controlled Vocabulary column	Required
Assay	Char	50	Assay Name	RNA -Seq, DNA-Seq, WGS, WES, DNA targeted (panels), PCR, qRT-PCR, qPCR, RT-PCR, ddPCR, HRM PCR, Atila, Xpert HPV	--	Must match one of the values listed in the Controlled Vocabulary column	Required
Lab	Char	25	Testing Laboratory Name	[Lab A, Lab B]	--	Must match one of the values listed in the Controlled Vocabulary column	Required
Lab Institution	Char	50	Testing Laboratory Institution	[Institution A, Institution B]	--	Must match one of the values listed in the Controlled Vocabulary column	Required

PI	Char	25	Testing Laboratory Principal Investigator	[Principal Investigator A, Principal Investigator B]	--	Must match one of the values listed in the Controlled Vocabulary column	Required for university-based laboratories
Specimen Type	Char	25	Specimen Type	PBMC, Serum, Plasma, Whole Blood, Bone Marrow, Fresh Tumor Tissue, Frozen Tumor Tissue, Anal Swab, Cervical Swab, Penile/Scrotal swab	--	Must match one of the values listed in the Controlled Vocabulary column	Required
Instrument	Char	50	Instrument	Illumina HiSeq4000, Illumina NovaSeq 6000, etc.	[provide reference]	Must match one of the values listed in the Controlled Vocabulary column	Required
Units	Char	2	Units of results, as applicable	Cycles, variant allele frequency, Log2Ratio, z-score, genome equivalents/mL genome copies / mL	--	Must match one of the values listed in the Controlled Vocabulary column	Expected for PCR Assays
Limit of Detection	Num	5,2	Lower limit of detection of assay	--	--	Format must be	Expected for

						xx.xx	PCR Assays
Lower Limit of Quantitation	Num	5,2	Lower limit of quantification of assay	--	--	Format must be xx.xx	Expected for PCR Assays
Upper Limit of Quantitation	Num	7,2	Upper limit of quantification of assay	--	--	Format must be xxxx.xx	Expected for PCR Assays
Controls	Char	50	Experimental controls	ERCC ExFold RNA Spike-In Mix (ThermoFisher)	--	Must match one of the values listed in the Controlled Vocabulary column	Expected
RNA Fractionization Method	Char	50	RNA Fractionization Method		--		Expected for RNA sequencing assays
RNA Species	Char	50	RNA Species	Total RNA, mRNA, rRNA, etc.	--		Expected for RNA sequencing assays
RNA extraction	Char	50	RNA extraction method	RNeasy Mini kit (Qiagen)	--		Expected for RNA sequencing assays
Nucleic Acid Extraction Method	Char	50	Nucleic Acid Extraction Method	RNeasy Mini kit (Qiagen)	--		Expected for RNA sequencing assays
Library kit	Char	50	Library kit	TruSeq mRNA Stranded Kit	--		Expected for sequencing assays
Targeted	Num	3	Targeted Read	--	--	Format must be	Expected for

Read length			length			xxx	sequencing assays
Targeted coverage [M reads]	Num	2,1	Targeted coverage [M reads]	--	--	Format must be xx.x	Expected for sequencing assays
Pairedness	Char	25	Pairedness	Single End, Paired End	--	Must match one of the values listed in the Controlled Vocabulary column	Expected for sequencing assays
Strandedness	Char	25	Strandedness	Unstranded	--	Must match one of the values listed in the Controlled Vocabulary column	Expected for sequencing assays
Cell Line	Char	25	Cell line used for assays measuring viral load		--	Must match one of the values listed in the Controlled Vocabulary column	Expected for viral sequencing assays

6. Quality Metrics

Sample quality metrics will be collected and reported in the following format. Requirements for collection are noted in the last column titled “Collection Requirement”. Sample quality metrics will be submitted in a .csv file to the ODMC.

- Filename will be [Protocol ID]_[assay]_[date]_quality_metrics.csv
- The data must have the columns listed exactly as in the “Variable Name” column in the table below. Included values must be included as listed in the “Validation Rules” column in the table below.

Table 8: Variables to be Submitted in [Protocol ID] [assay] [date] quality_metrics.csv

Variable Name	Type	Length	Description	Reference	Validation Rules	Collection Requirement
Protocol	Num	5	Protocol ID	--	Must match ODMC-specified value for the study	Required
Specimen ID	Num	9	Specimen number, refers to the GlobalTrace barcode number printed on each cryovial	N/A	Must be 9 characters long. If a Specimen is being entered as a reference (control) sample, the SN will be labeled as [specified value]	Required
% Q30 score	Num	5,2	Percentage of bases with sequencing quality score of at least 30	https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html	Must be documented as a percentage, 0 – 99.99 %.	Expected for sequencing assays
Mean depth	Num	9	The sum of the mapped read depths at each reference base position, divided by the number of known bases in the reference	https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html	Format must be xx	Expected for sequencing assays
Median depth	Num	9	The median mapped read depth	https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html	Format must be xx	Expected for sequencing assays
Mean insert size	Num	9	Mean length of the DNA (or RNA) being sequenced	https://en.wikipedia.org/wiki/Insert_(molecular_biology)	Format must be xxx	Expected for sequencing assays

Variable Name	Type	Length	Description	Reference	Validation Rules	Collection Requirement
Median insert size	Num	9	Median length of the DNA (or RNA) being sequenced	https://en.wikipedia.org/wiki/Insert_(molecular_biology)	Format must be xxx	Expected for sequencing assays
Cell viability (%)	Num	5,2	Percent cell viability prior to testing	N/A	Must be documented as a percentage, 0 – 99.99 %.	Expected for assays performed on PBMCs
Cell count	Num	9	Number of cells measured prior to testing	N/A	N/A	Optional for assays performed on PBMCs
Total RNA [ng]	Num	2	Total RNA amount used for assay	N/A	Format must be xx	Expected for RNA sequencing
Total DNA [ng]	Num	2	Total DNA amount used for assay	N/A	Format must be xx	Expected for MRD sequencing
RIN	Num	2	RNA integrity number	https://en.wikipedia.org/wiki/RNA_integrity_number RNA quality will be assessed using an Agilent Bioanalyzer.	Format must be xx	Expected for RNA sequencing
DV200	Num	2	Percentage of RNA fragments > 200 nucleotides	https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/evaluating-rna-quality-from-ffpe-samples-technical-note-470-2014-001.pdf	Format must be xx. Value reported in %	Expected for RNA sequencing from FFPE samples

Variable Name	Type	Length	Description	Reference	Validation Rules	Collection Requirement
Freeze/Thaw Cycles	Num	1	Number of freeze/thaw cycles prior to testing	N/A	N/A	Expected

7. Manifest of File Collection to be Submitted to the ODMC

The manifest of the file collection describes the collection of files that are expected along with the transferred files to ensure the integrity of the collection. The following files are expected to be listed in the manifest, at minimum: [Protocol ID]_[assay]_[date]_workflow.csv, [Protocol ID]_[assay]_[date]_exp_data.csv, [Protocol ID]_[assay]_[date]_specimen_linkage.csv, text file with experimental summary, [Protocol ID]_[assay]_[date]_exp_metadata.csv, [Protocol ID]_[assay]_[date]_quality_metrics.csv, [Protocol ID]_[assay]_[date]_exp_manifest.csv.

Manifests will be submitted in a .csv to the ODMC. Requirements for collection are noted in the last column titled “Collection Requirement”.

- Filename will be [Protocol ID]_[assay]_[date]_exp_manifest.csv
- The data must have the columns listed exactly as in the “Variable Name” column. Included values must be included as listed in the “Controlled Vocabulary” and “Validation Rules” columns in the table below.

Table 9: Variables to be Submitted in [Protocol ID] [assay] [date] exp manifest.csv

Variable Name	Type	Length	Description	Controlled Vocabulary	Reference	Validation Rules	Collection Requirement
Protocol	Num	5	Protocol ID	--	N/A	Must match ODMC-specified value for the study	Required
File Name	Char	256	Name of the file	--	N/A	--	Required
File Type	Char	256	Type of the file	Raw, Interim-Processed, Final-Processed, Meta-Data, Quality-Metrics	N/A	Must match one of the values listed in the Controlled Vocabulary column	Required
Sha256Sum	Char	256	Sha256 Checksum	--	See Data Preparation Instructions available at https://amcoperations.com/committee/amc-data-commons	--	Required
Timestamp	yyyy-mm-dd hh:mm		Data / Time of file creation	--	N/A	--	Required

Assay results will be reviewed for accuracy and compliance with specified data standards by the laboratory staff prior to transfer per the laboratory's procedures and SOPs.

8. ODMC Data Processing

No data processing is expected to be completed by the ODMC.

9. Approval

Data Standardization Plan Approval	
By signing this data transfer plan, I approve of the included specifications.	
Laboratory Resource Committee Chair:	
Electronically signed by: Jeffrey Bethony Reason: I am Approving this document Date: Oct 14, 2026 16:31:06 ADT Signature: <u>Jeffrey Bethony</u> Jeffrey Bethony 10/14/2026 Signature Date	